

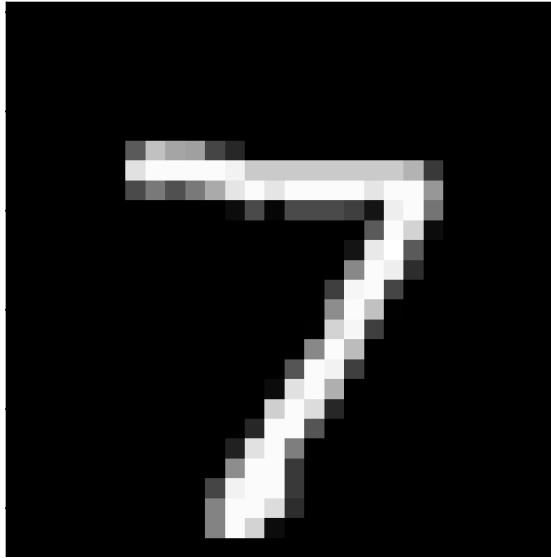
# Workflow

Ki Hyun Kim

[nlp.with.deep.learning@gmail.com](mailto:nlp.with.deep.learning@gmail.com)

# Our Objective

- 주어진 데이터에 대해서 결과를 내는 가상의 함수를 **모사**하는 함수를 만드는 것
  - e.g. 주어진 숫자 그림을 보고 숫자 맞추기



# Working Process

## 문제 정의

- 단계를 나누고 simplify
- $x$ 와  $y$ 를 정의



## 데이터 수집

- 문제 정의에 따른 수집
- 필요에 따라 레이블링



## 데이터 전처리 및 분석

- 형태를 가공
- 필요에 따라 EDA 수행



## 알고리즘 적용

- 가설을 세우고 구현/적용



## 평가

- 실험 설계
- 테스트셋 구성



## 배포

- RESTful API를 통한 배포
- 상황에 따라 유지/보수

# 문제 정의

- 가장 중요한 부분
- 풀고자 하는 문제를 단계별로 나누고 simplify 하여야 한다.
  - e.g. 라면 끓이기
- 신경망이라는 함수에 넣기 위한  $x$ 와 결과값  $y$ 가 정의 되어야 한다.

$$y = f(x)$$

# 데이터 수집

- 문제 정의에 따라 정해진  $x$ 와  $y$ .
- 풀고자 하는 문제의 영역에 따라 수집 방법이 다름
  - 자연어처리, 컴퓨터비전: crawling
  - 데이터분석: 실제 수집한 데이터
- 필요에 따라 레이블링(labeling) 작업을 수행
  - 자동적으로 레이블(label)이  $y$ 로 주어질 수도 있음
  - 하지만 대부분의 경우, 레이블이 따로 필요함
  - 비지도학습(unsupervised learning)을 기대하지 말자

# 데이터 전처리 및 분석

- 수집된 데이터를 신경망에 넣어주기 위한 형태로 가공하는 과정
  - 입출력 값을 정제(cleaning & normalization)
- 이 과정에서 탐험적분석(Exploratory Data Analysis, EDA)이 필요
  - 데이터에 알맞는 알고리즘을 찾기 위함
  - 자연어처리, 컴퓨터비전의 경우에는 생략되기도 함
- 영상처리(computer vision)의 경우, 데이터 증강(augmentation)이 수행됨
  - e.g. rotation, flipping, shifting 등의 간단한 연산

# 알고리즘 적용

- 데이터에 대해 가설을 세우고, 해당 가설을 위한 알고리즘(모델)을 적용

# 평가

- 문제 정의에 따른 공정하고 올바른 평가 방법 필요
  - 가설을 검증하기 위한 실험 설계
- 테스트 셋(test set) 구성
  - 너무 쉽거나 어렵다면 판별력이 떨어짐
  - 실제 데이터와 가장 가깝게 구성되어야 함
- 정량적 평가(extrinsic evaluation)와 정성적 평가(intrinsic evaluation)로 나뉨



# 배포

- 학습과 평가가 완료된 모델 weights 파일을 배포함
- RESTful API 등을 통해 wrapping 후 배포
- 데이터 분포의 변화에 따른 모델 업데이트 및 유지/보수가 필요할 수 있음